Automatic detection of the foveal center in optical coherence tomography

Bart Liefers,[1,2,*] Freerk G. Venhuizen,[1,2] Vivian Schreur,[2] Bram van Ginneken,[1] Carel Hoyng,[2] Sascha Fauser,[3,4] Thomas Theelen,[1,2] and Clara I. Sánchez[1,2]

[1]*Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, the Netherlands*
[2]*Department of Ophthalmology, Radboud University Medical Center, Nijmegen, the Netherlands*
[3]*Roche Pharma Research and Early Development, F. Hoffmann-La Roche Ltd, Basel, Switzerland*
[4]*Cologne University Eye Clinic, Cologne, Germany*
[*]*Bart.Liefers@radboudumc.nl*

Abstract: We propose a method for automatic detection of the foveal center in optical coherence tomography (OCT). The method is based on a pixel-wise classification of all pixels in an OCT volume using a fully convolutional neural network (CNN) with dilated convolution filters. The CNN-architecture contains anisotropic dilated filters and a shortcut connection and has been trained using a dynamic training procedure where the network identifies its own relevant training samples. The performance of the proposed method is evaluated on a data set of 400 OCT scans of patients affected by age-related macular degeneration (AMD) at different severity levels. For 391 scans (97.75%) the method identified the foveal center with a distance to a human reference less than 750 µm, with a mean (± SD) distance of 71 µm ± 107 µm. Two independent observers also annotated the foveal center, with a mean distance to the reference of 57 µm ± 84 µm and 56 µm ± 80 µm, respectively. Furthermore, we evaluate variations to the proposed network architecture and training procedure, providing insight in the characteristics that led to the demonstrated performance of the proposed method.

OCIS codes: (110.4500) Optical coherence tomography; (100.4996) Pattern recognition, neural networks; (100.2960) Image analysis; (170.4470) Clinical applications; (170.4470) Ophthalmology.

References and links

1. P. A. Keane, S. Liakopoulos, K. T. Chang, M. Wang, L. Dustin, A. C. Walsh, and S. R. Sadda, "Relationship between optical coherence tomography retinal parameters and visual acuity in neovascular age-related macular degeneration," Ophthalmology **115**, 2206–2214 (2008).
2. S. M. Waldstein, A. Philip, R. Leitner, C. Simader, G. Langs, B. S. Gerendas, and U. Schmidt-Erfurth, "Correlation of 3-dimensionally quantified intraretinal and subretinal fluid with visual acuity in neovascular age-related macular degeneration," JAMA Ophthalmology **134**, 182–190 (2016).
3. Diabetic Retinopathy Clinical Research Network, "Relationship between optical coherence tomography–measured central retinal thickness and visual acuity in diabetic macular edema," Ophthalmology **114**, 525–536 (2007).
4. T. Otani, Y. Yamaguchi, and S. Kishi, "Correlation between visual acuity and foveal microstructural changes in diabetic macular edema," Retina **30**, 774–780 (2010).
5. M. Adhi and J. S. Duker, "Optical coherence tomography–current and future applications," Curr. Opin. Ophthalmol. **24**, 213 (2013).
6. W. Geitzenauer, C. K. Hitzenberger, and U. M. Schmidt-Erfurth, "Retinal optical coherence tomography: past, present and future perspectives," Br. J. Ophthalmol. (2010).
7. M. R. Hee, J. A. Izatt, E. A. Swanson, D. Huang, J. S. Schuman, C. P. Lin, C. A. Puliafito, and J. G. Fujimoto, "Optical coherence tomography of the human retina," Arch. Ophthalmol. **113**, 325–332 (1995).
8. M. Niemeijer, M. D. Abràmoff, and B. van Ginneken, "Fast detection of the optic disc and fovea in color fundus photographs," Med. Image Anal. **13**, 859–870 (2009).
9. A. S. Maheshwary, S. F. Oster, R. M. Yuson, L. Cheng, F. Mojana, and W. R. Freeman, "The association between percent disruption of the photoreceptor inner segment–outer segment junction and visual acuity in diabetic macular edema," Am. J. Ophthalmol. **150**, 63–67 (2010).
10. C. Balaratnasingam, M. Inoue, S. Ahn, J. McCann, E. Dhrami-Gavazi, L. A. Yannuzzi, and K. B. Freund, "Visual acuity is correlated with the area of the foveal avascular zone in diabetic retinopathy and retinal vein occlusion," Ophthalmology **123**, 2352–2367 (2016).

11. P. C. Issa, M. C. Gillies, E. Y. Chew, A. C. Bird, T. F. Heeren, T. Peto, F. G. Holz, and H. P. Scholl, "Macular telangiectasia type 2," Progress in Retinal and Eye Research **34**, 49–77 (2013).
12. A. Chan, J. S. Duker, T. H. Ko, J. G. Fujimoto, and J. S. Schuman, "Normal macular thickness measurements in healthy eyes using stratus optical coherence tomography," Arch. Ophthalmol. **124**, 193–198 (2006).
13. P. A. Campochiaro, J. S. Heier, L. Feiner, S. Gray, N. Saroj, A. C. Rundle, W. Y. Murahashi, R. G. Rubio, BRAVO Investigators, "Ranibizumab for macular edema following branch retinal vein occlusion: six-month primary end point results of a phase iii study," Ophthalmology **117**, 1102–1112 (2010).
14. P. Massin, A. Erginay, B. Haouchine, A. B. Mehidi, M. Paques, and A. Gaudric, "Retinal thickness in healthy and diabetic subjects measured using optical coherence tomography mapping software." Eur. J. Ophthalmol. **12**, 102–108 (2001).
15. Age-Related Eye Disease Study Research Group, "A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins c and e, beta carotene, and zinc for age-related macular degeneration and vision loss: Areds report no. 8," Arch. Ophthalmol. **119**, 1417–1436 (2001).
16. C. D. Regillo, D. M. Brown, P. Abraham, H. Yue, T. Ianchulev, S. Schneider, N. Shams, "Randomized, double-masked, sham-controlled trial of ranibizumab for neovascular age-related macular degeneration: Pier study year 1," Am. J. of Ophthalmol. **145**, 239–248 (2008).
17. B. Gerendas, S. Waldstein, J. Lammer, A. Montuoro, G. Bota, C. Simader, U. Schmidt-Erfurth, and Vienna Reading Center, "Centerpoint replotting and its effects on central retinal thickness in four prevalent SD-OCT devices," Invest. Ophthalmol. Vis. Sci. **53**, 4114–4114 (2012).
18. F. Wang, G. Gregori, P. J. Rosenfeld, B. J. Lujan, M. K. Durbin, and H. Bagherinia, "Automated detection of the foveal center improves SD-OCT measurements of central retinal thickness," Ophthalmic Surgery, Lasers and Imaging Retina **43**, S32–S37 (2012).
19. S. Tick, F. Rossant, I. Ghorbel, A. Gaudric, J.-A. Sahel, P. Chaumet-Riffaud, and M. Paques, "Foveal shape and structure in a normal population," Invest. Ophthalmol. Vis. Sci. **52**, 5105–5110 (2011).
20. J. Wu, S. M. Waldstein, A. Montuoro, B. S. Gerendas, G. Langs, and U. Schmidt-Erfurth, "Automated fovea detection in spectral domain optical coherence tomography scans of exudative macular disease," Int. J. of Biomed. Imaging **2016** (2016).
21. S. J. Chiu, X. T. Li, P. Nicholas, C. A. Toth, J. A. Izatt, and S. Farsiu, "Automatic segmentation of seven retinal layers in SDOCT images congruent with expert manual segmentation," Opt. Express **18**, 19413–19428 (2010).
22. A. Lang, A. Carass, M. Hauser, E. S. Sotirchos, P. A. Calabresi, H. S. Ying, and J. L. Prince, "Retinal layer segmentation of macular OCT images using boundary classification," Biomed. Opt. Express **4**, 1133–1152 (2013).
23. R. Kafieh, H. Rabbani, F. Hajizadeh, M. D. Abramoff, and M. Sonka, "Thickness mapping of eleven retinal layers segmented using the diffusion maps method in normal eyes," J. Ophthalmol. **2015** (2015).
24. S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema," Biomed. Opt. Express **6**, 1172–1194 (2015).
25. A. Montuoro, S. M. Waldstein, B. S. Gerendas, U. Schmidt-Erfurth, and H. Bogunović, "Joint retinal layer and fluid segmentation in OCT scans of eyes with severe macular edema using unsupervised representation and auto-context," Biomed. Opt. Express **8**, 1874–1888 (2017).
26. G. Litjens, T. Kooi, B. E, Bejnordi, A. A. A. Setio, F. Ciompi M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Med. Image Anal. **42**, 60–88 (2017).
27. L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu, "Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search," Biomed. Opt. Express **8**, 2732–2744 (2017).
28. F. G. Venhuizen, B. van Ginneken, B. Liefers, M. J. van Grinsven, S. Fauser, C. Hoyng, T. Theelen, and C. I. Sánchez, "Robust total retina thickness segmentation in optical coherence tomography images using convolutional neural networks," Biomed. Opt. Express **8**, 3292–3316 (2017).
29. S. P. K. Karri, D. Chakraborty, and J. Chatterjee. "Transfer Learning Based Classification of Optical Coherence Tomography Images with Diabetic Macular Edema and Dry Age-Related Macular Degeneration," Biomed. Opt. Express **8**, 579–592 (2017).
30. A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, "ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," Biomed. Opt. Express **8**, 3627–3642 (2017).
31. B. Liefers, F. G. Venhuizen, T. Theelen, C. Hoyng, B. van Ginneken, and C. I. Sánchez, "Fovea detection in optical coherence tomography using convolutional neural networks," Proc. SPIE 10133, (2017).
32. J. P. van de Ven, D. Smailhodzic, C. J. Boon, S. Fauser, J. M. Groenewoud, N. V. Chong, C. B. Hoyng, B. J. Klevering, and A. I. den Hollander, "Association analysis of genetic and environmental risk factors in the cuticular drusen subtype of age-related macular degeneration," Molecular Vision **18**, 2271–2278 (2012).
33. S. Fauser, D. Smailhodzic, A. Caramoy, J. P. H. van de Ven, B. Kirchhof, C. B. Hoyng, B. Jeroen Klevering, S. Liakopoulos, and A. I. den Hollander, "Evaluation of serum lipid concentrations and genetic variants at high-density lipoprotein metabolism loci and TIMP3 in age-related macular degeneration," Invest. Ophthalmol. Vis. Sci. **52**, 5525–5528 (2011).

34. S. Farsiu, S. J. Chiu, R. V. O'Connell, F. A. Folgar, E. Yuan, J. A. Izatt, C. A. Toth, AREDS 2 Ancillary Spectral Domain Optical Coherence Tomography Study Group, "Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography," Ophthalmology **121**, 162–172 (2014).
35. F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv preprint arXiv:1511.07122 (2015).
36. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," arXiv preprint arXiv:1606.00915 (2016).
37. M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in "Wavelets," (Springer, 1990), pp. 286–297.
38. Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," arXiv e-prints **abs/1605.02688** (2016).
39. S. Dieleman, J. Schlüter, C. Raffel, E. Olson, S. K. Sønderby, D. Nouri, "Lasagne: First release." http://dx.doi.org/10.5281/zenodo.27878 (2015).
40. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in "IEEE Conf. Comput. Vis. Pattern Recognit.", (2015), pp. 3431–3440.
41. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv preprint arXiv:1512.03385 (2015).
42. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," MICCAI 2015: 18th International Conference 9351, 234âĂŞ241 (2015).
43. A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in "Proceedings of the International Machine Learning Society", **30** (2013)
44. M. J. van Grinsven, B. van Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sánchez, "Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images," IEEE Trans. Med. Imag. **35**, 1273–1284 (2016).
45. M. J. Hogan and J. JA Weddell, "Histology of the human eye: an atlas and textbook," (1971).
46. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556 (2014).
47. A. Govetto, R. A. Lalane III, D. Sarraf, M. S. Figueroa, and J. P. Hubschman, "Insights into epiretinal membranes: Presence of ectopic inner foveal layers and a new optical coherence tomography staging scheme," Am. J. Ophthalmol. **175**, 99 – 113 (2017).
48. L. Fang, S. Li, R. P. McNabb, Q. Nie, A. N. Kuo, C. A. Toth, J. A. Izatt and S. Farsiu "Fast acquisition and reconstruction of optical coherence tomography images via sparse representation," IEEE Trans. Med. Imag. **32**, 2034–2049, (2013).
49. L. Fang, S. Li, D. Cunefare, and S. Farsiu, "Segmentation based sparse reconstruction of optical coherence tomography images," IEEE Trans. Med. Imag. **36**, 407–421 (2017).

1. Introduction

The fovea is a region located near the center of the retina with the highest concentration of cones, photoreceptor cells responsible for color vision. As a result of this elevated concentration, the fovea is responsible for central vision and high spatial acuity. Consequently, any small alteration of its morphology or the presence of abnormalities in this area directly affects visual acuity [1–4]. Closely monitoring the fovea and its changes is therefore highly important for the prevention and assessment of vision threatening conditions.

Optical coherence tomography (OCT) is a non-invasive imaging technology that allows a detailed in-vivo analysis of the interior of the retina and, particularly, the fovea. This technique is based on low-coherence interferometry, where differences in back-scattering properties reveal the layered structure of the retina and produce high resolution images of cross sections of the retina. The resolution and image quality of OCT scans have improved rapidly in recent years, making it a leading imaging modality in clinical practice [5, 6]. With OCT a more reliable estimation of the exact location of the fovea can be made compared to en-face modalities such as color fundus imaging, especially in retinal pathology [7, 8].

Studying the morphology of the fovea and its surroundings as seen in OCT helps in the early detection and understanding of retinal diseases, such as age-related macular degeneration (AMD) [1, 2], diabetic macular edema (DME) [4, 9], retinal vein occlusion (RVO) [10] or macular telangiectasia [11]. Additionally, the foveal position in OCT scans is a key reference landmark for the extraction of reliable quantitative biomarkers. Central macular thickness (CMT), defined

as the average retinal thickness in the 1 mm disk centered on the fovea [12], and markers based on fovea-centered measurement grids, such as the Early Treatment Diabetic Retinopathy Study (ETDRS) grid, have become important quantitative measurements to monitor disease progression and treatment response objectively [13, 14]. These values are commonly used as biomarkers in large population studies and clinical trials [15, 16]. However, these measurements highly depend on a precise location of the foveal center, as deviations from the correct position have a detrimental impact in their calculated values and consequently their reliability [17, 18].

During OCT acquisition, the geometric center of the OCT volume is usually placed at the foveal center. However, this center often does not align with the actual foveal center due to the presence of pathology or poor participant's fixation [18]. Consequently, the scan center might not be a reliable estimation of the correct foveal center. Manual corrections are therefore required to prevent misleading outcomes, which is time consuming, undesirable and even unfeasible in large data sets. Fully automated methods are then required for the accurate and efficient localization of the foveal center in OCT scans.

In a healthy retina, the fovea in OCT is characterized by 1) a concave dip in the retina, 2) the absence of the inner layers of the retina and 3) a slight thickening of the ellipsoid zone (EZ) [19]. In pathological retinas, the presence of abnormalities can drastically alter the expected appearance of the fovea, making the detection of the fovea a challenging problem, even for experienced human graders [18, 20]. Figure 1 shows an example of a healthy and a pathological fovea.

The thinning and confluence of retinal layers near the fovea has been used by several authors to estimate the fovea position, which was required to improve results of layer segmentation algorithms [21–24]. In [25] the position of the fovea is estimated by training a random forest regressor based on thickness estimates to predict the distance to the fovea for each A-scan. By using a random sample consensus, the A-scan with the final fovea position is found. However, disrupting retinal structures, such as fluid or atrophy, may render the deduction of the foveal position from thickness measurements unreliable. In [20] automated detection of the fovea in pathological retinas was approached by generalizing the fovea morphology into three fixed categories of foveal shape: normal, minor and absent foveal depression. This limited number of foveal shapes still prevents the method to account for the large variation in foveal morphology shown in OCT.

In recent years the use of convolutional neural networks (CNNs) has gained popularity in medical image analysis [26]. Their application to OCT has been demonstrated by e.g. retinal layer segmentation [27, 28], identification of retinal pathologies [29] or segmentation of fluid in combination with retinal layers [30]. In this paper, we propose a method based on CNNs for the automated detection of the foveal center in OCT volumes, building upon our previous work described in [31]. We introduce a novel fully CNN architecture which combines anisotropic dilated convolutions and a shortcut connection. This architecture provides a large contextual window to account for the wide variability of foveal shapes, while maintaining the ability to make predictions at pixel level accuracy. In contrast to previously proposed approaches, the method is morphology-agnostic, avoiding a priori assumptions of the foveal shape or preprocessing steps such as layer segmentation, which makes it robust to the presence of various disrupting abnormalities. The performance of our method is evaluated on OCT volumes from patients affected by AMD at different severity levels. An expert human grader and two independent human observers manually annotated the foveal center, providing a reference location and an estimate of human performance for this task. We examine variations to both the proposed network architecture and the training procedure, to gain insight in the characteristics of the proposed method that are essential for the performance. Furthermore, we demonstrate how the proposed method could be applied in a more general context, for retinal pathologies other than AMD, and independent of the vendor or the scanning protocol that was used.

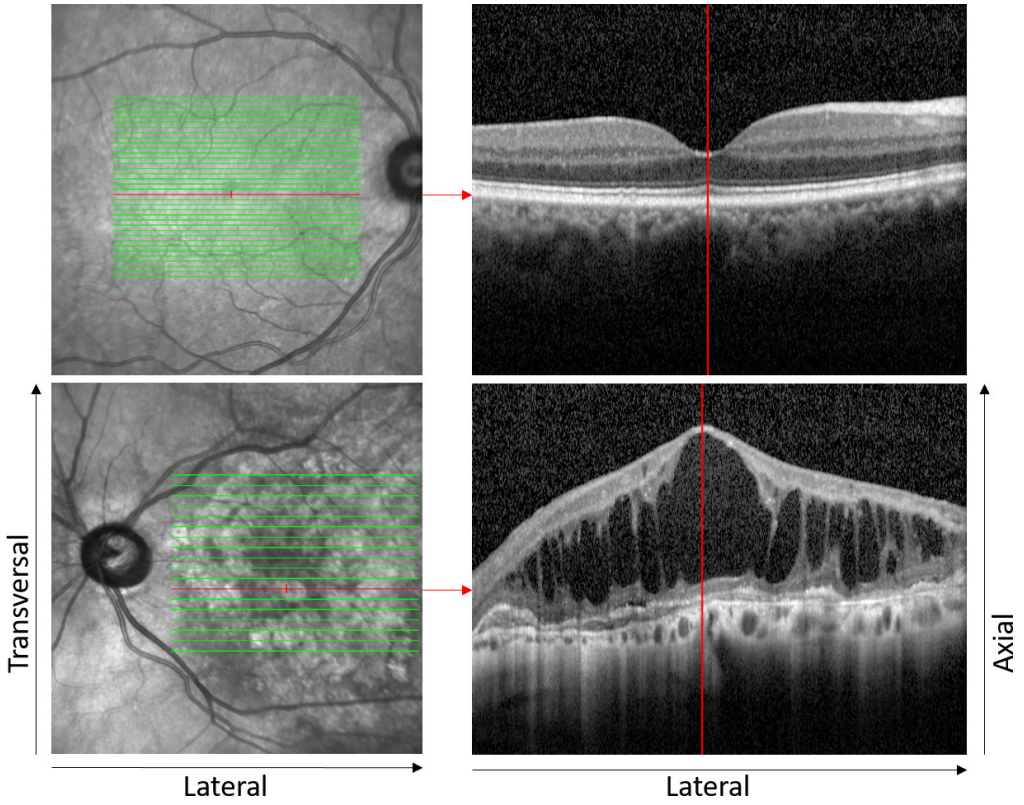

Fig. 1. Examples of a fovea as seen in scanning laser ophthalmoscopy (SLO, left column) and OCT (right column) in a healthy retina (top) and a pathological retina (bottom) where the presence of cysts affects the expected appearance of the fovea. The green lines on the SLO images show the position of the B-scans. The red line on the SLO indicates the B-scan with the foveal center (as annotated by the reference observer). The red line on the OCT indicates the lateral position of the fovea. Transversal and lateral are used to indicate directions in the en-face plane. Axial refers to the depth of the B-scan in the retina. The top image has been acquired with a dense scanning protocol (37 B-scans with a transversal resolution of 119 µm). The bottom image has been acquired with a sparse scanning protocol (19 B-scans with a transversal resolution of 257 µm). The images are cropped in the axial direction for visualization purposes.

2. Data

A total of 2,244 OCT volumes from the European Genetic Database (EUGENDA), a large multi-center database for clinical and molecular analysis of AMD [32], were used for the development and evaluation of this study. The OCT volumes were acquired using a Heidelberg Spectralis HRA+OCT (Heidelberg Engineering, Heidelberg, Germany) at a wavelength of 870 nm and consisted of a stack of horizontal slices (B-scans) at different resolutions. The lateral resolution ranges from 6 µm to 14 µm. The axial resolution is 3.9 µm. The transversal resolution (distance between B-scans) varies, but the vast majority of scans is acquired at a transversal resolution of either ~120 µm (37 B-scans) or ~240 µm (19 B-scans). Figure 1 shows an example of the dense scanning protocol (37 B-scans) and the sparse scanning protocol (19 B-scans). For a subset of OCT volumes the AMD severity level is graded, based on the assessment of a color fundus image acquired at the same time, following the Cologne Image Reading Center and Laboratory

(CIRCL) grading protocol [33]. This grading protocol includes the following AMD severity levels: no AMD, early AMD, intermediate AMD, advanced AMD with choroidal neovascularization (CNV), and advanced AMD with geographic atrophy (GA). The EUGENDA study was performed according to the tenets set forth in the Declaration of Helsinki, and approved by the Institutional Review Board. Written informed consent was obtained before enrolling patients in EUGENDA.

A development set of 1844 OCT volumes from 399 subjects was selected and split into a **training set** of 1744 volumes and a **validation set** of 100 volumes for algorithm training and optimization, respectively. An independent **test set** of 400 OCT volumes from 400 eyes of 238 subjects was used for the evaluation of the algorithm performance. The test volumes were randomly and evenly extracted from different AMD severity levels, with 100 volumes for each of the four AMD severity levels, i.e., no AMD, early AMD, intermediate AMD and advanced AMD. Scans with large registration errors or poor image quality were excluded.

Before processing, all B-scans have been resampled to a constant resolution of $11.5\,\mu\text{m} \times 3.9\,\mu\text{m}$ (lateral $\times$ axial). The lateral width of the scans in the test set varies from 391 pixels (4.5 mm) to 888 pixels (10.2 mm) with axial depth of 496 pixels (1.9 mm).

For all selected scans in the development set, the location of the foveal center was manually annotated as a single 3D coordinate in the OCT volume by a human grader. Only OCT information was used for the manual localization of the fovea, although additional imaging information, such as the corresponding infrared scanning laser ophthalmoscopy (SLO) image, might be available. The human grader was instructed to indicate the point on the inner limiting membrane (ILM) closest to the foveal center, identified by the aforementioned characteristics of the fovea. In case of unnatural foveal deformations due to pathology, the grader had to indicate the point of minimal thickness of the retinal nerve fibre layer. For all OCT volumes in the test set, an experienced grader (VS) and two independent graders (BL, CS) annotated the foveal center following the same protocol as in the development set. The annotations from the expert grader were considered the reference standard for this study, while the annotations from the other two observers were collected for comparative analysis of human performance.

To validate whether the method would generalize well to other retinal pathologies, a set of 50 OCT volumes from patients affected by diabetic macular edema (DME) and 50 OCT volumes from patients affected by central serous chorioretinopathy (CSC) was used. For this set the reference fovea was set by a single observer (BL). All OCT volumes in this set were collected using a Heidelberg Spectralis OCT scanner.

Additionally, an **external set** was used in order to evaluate how the proposed method can be adapted to be used on data with different properties. This publicly available data set consists of 384 volumes acquired with a Bioptigen scanner [34]. Compared to the scans from Heidelberg Spectralis, these scans have a higher transversal resolution of 67 µm per pixel (100 B-scans per volume), but they present a higher level of noise within each B-scan. The scans in this set have been graded as control (115 scans) or intermediate AMD (269 scans). For this data set manual delineations of retinal layers are available for the central 5 mm of the retina. The center of these annotations coincides with the foveal center, and is hence chosen as the reference location for this data set. A subset of 40 images (20 control, 20 AMD) has been randomly selected for development. The remaining 344 images are used for evaluation.

3. Methods

In the proposed approach, the identification of the fovea is addressed as a classification problem, where every pixel in every B-scan is classified using a CNN as either being the true fovea center or background, based on a contextual window of information around the pixel. The pixel with the maximum likelihood after classification will then indicate the location of the fovea. However, this approach faces the following challenges: 1) for an accurate classification, the contextual window (defined by the receptive field of the network) should be large enough to include sufficient

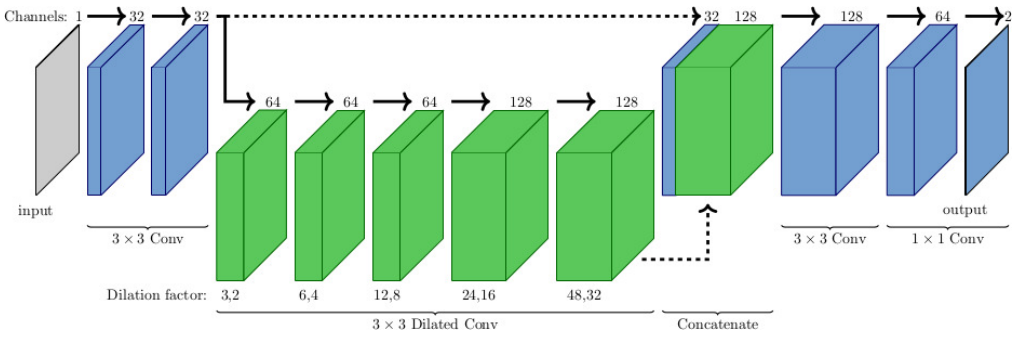

Fig. 2. Visualization of the network architecture. The solid arrows represent convolution operations. The dashed arrows represent a copy operation. The blue layers represent the output of the regular convolutions. The green layers represent the output of the dilated convolutions. The resolution of the feature maps remains the same throughout the network. Because the network is fully convolutional, it can be applied to inputs of arbitrary size.

structural aspects, such as the confluence of retinal layers, while maintaining pixel precision and computational requirements; 2) for effective training of the CNN, the extreme class-imbalance of the classification problem (the few pixels representing the foveal center versus the large amount of background pixels that only rarely contain relevant information) should be considered. To address the first challenge, we propose a novel CNN architecture containing anisotropic dilated convolutions [35] (also known as atrous convolutions [36, 37]) enclosed by a shortcut connection. Compared to regular convolutions, dilated convolutions allow for a more efficient expansion of the receptive field of the network (and consequently the contextual window), while maintaining tractable computational requirements and pixel accuracy. The second challenge is addressed by a dynamic training procedure where the provisional classifications of the CNN are used to identify relevant training samples. Details about the proposed architecture, depicted in Fig. 2, and training procedure are provided in the following sections. The proposed method is implemented in Python 2.7 using the Theano [38] and Lasagne [39] packages.

3.1. Network architecture

The proposed architecture makes use of a fully CNN architecture [40]. In contrast to traditional CNNs, which usually include one or more fully connected layers, fully CNNs exclusively make use of convolutional layers (optionally also pooling layers). Because spatial correlation between input and output is conserved in fully CNNs, they can be applied efficiently to inputs of arbitrary size. Within a fully CNN, the receptive field of the network is gradually expanded by stacking multiple convolutional layers. A regular 3×3 convolutional layer increases the receptive field by adding a border of one pixel around the output of the previous layer, and can in this way be used for linear expansion of the receptive field. In order to obtain a large receptive field, a large number of convolutional layer need to be stacked, which can have a detrimental impact on performance [41] and is memory-wise infeasible. Pooling layers, in contrast, are very effective in expanding the receptive field. A 2×2 pooling doubles the receptive field of the previous layer, and therefore allows for an exponential expansion of the receptive field. However, after every pooling operation the resolution of the output, and consequently the pixel level accuracy, is decreased. Although the lost resolution can be recovered using interpolation techniques such as shift-and-stitch [40], this is computationally prohibitive. Dilated convolutions, like pooling operations, can be used to obtain exponential expansion of the receptive field by incorporating a specific spacing between the parameters of the filter [35], as shown in Fig. 3. However, unlike pooling, they preserve the resolution of the input image. In this way, a fast expansion of the

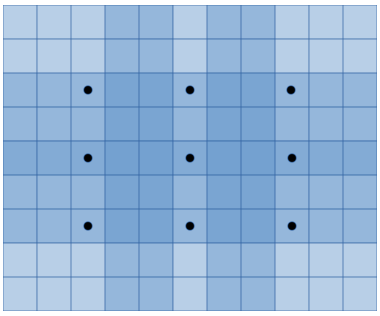

Fig. 3. Visualization of the dilated convolution filter at layer 3, with a receptive field of 11 × 9 pixels. The pixels with black dots are included in this filter. These pixels have a receptive field of 5 × 5 pixels each, as a result of the previous two convolution layers. The different shades in the figure represent the overlap of these subfields.

receptive field can be achieved with just a few layers and without losing resolution.

The proposed architecture incorporates five layers with dilated convolutions in order to increase the receptive field R. The receptive field of a CNN with dilated convolution filters can be calculated as:

$$R_i = R_{i-1} + d_i \cdot (k_i - 1) \tag{1}$$

where R_i is the receptive field at layer i, d_i is the size of the dilation (number of pixels between the weights in the filter), and k_i is the size of the filter. See Fig. 3 for a visualization of the dilated convolution filter and the receptive field at Layer 3. By increasing the dilation factor d_i for deeper layers a fast expansion of the receptive field can be obtained.

In the proposed network architecture shown in Fig. 2 and summarized in Table 1, the first two layers are regular 3×3 convolutions that provide low level feature extraction. Next, five layers with dilated convolutions are included to provide a large contextual window. Similar to [35], the size of the dilation factor is doubled for every layer deeper in the network. We choose anisotropic dilated convolutions, with a larger dilation factor in the lateral direction, because of the anisotropic resolution of the B-scans and taking into account that important structural clues, such as the confluence in retinal layers, are especially observed in the lateral direction. A shortcut connection combines the output of the last regular convolution (Layer 2) with the output of the last dilated convolution (Layer 7) by concatenating them in the feature dimension. Such a shortcut connection can combine information at different levels of abstraction or different scales and may increase accuracy and training efficiency [40–42]. Finally, a 3×3 convolution is applied and the last two layers perform feature combination using 1x1 convolutions. As a consequence, the network has a receptive field of 193×131 pixels (see Table 1). After every convolution layer (except the output layer), a leaky rectify non-linearity with leakiness 0.01 is applied [43]. At the output layer a softmax function is applied over the two classes, fovea and background.

3.2. Training procedure

The network is trained using standard backpropagation with RMSProp updates, and cross entropy as cost function. Every iteration a batch of 64 patches (32 fovea and 32 background) is supplied to the network. The size of the patches is 197×135 pixels, which yields an output of 5×5 pixels (due to the loss of border pixels after every convolution operation). Hence, by increasing the patch size from the minimal 193×131 pixels to 197×135, the error can be averaged over 25 pixels per patch. The fovea patches are always centered on the manually created reference location, while background patches are drawn from the remaining pixels in the B-scan. In order to avoid extracting background patches that are ambiguous, pixels in a region of 50×50 pixels

Table 1. Summary of the network architecture. The dilation (d_i) refers to the spacing in the dilated filters in the horizontal and vertical direction, respectively. The receptive fields indicate the size of the contextual window of all pixels that can influence the network output at that layer.

Layer (i)	Filter size (k_i)	Type	Dilation (d_i)	Channels	Receptive Field (R)
1	3×3	Conv	(1, 1)	32	3×3
2	3×3	Conv	(1, 1)	32	5×5
3	3×3	Dilated	(3, 2)	64	11×9
4	3×3	Dilated	(6, 4)	64	23×17
5	3×3	Dilated	(12, 8)	64	47×33
6	3×3	Dilated	(24, 16)	128	95×65
7	3×3	Dilated	(48, 32)	128	191×129
8	-	Concat (2 + 7)	-	32 + 128	191×129
9	3×3	Conv	(1, 1)	128	193×131
10	1×1	Conv	(1, 1)	64	193×131
11	1×1	Conv	(1, 1)	2	193×131

around the fovea are not included as background. Background patches are very abundant (up to 10^7 pixels per OCT volume), but only rarely very challenging: everything above or below the retina is almost completely black, and patches inside the retina are usually easily discriminated from the fovea. Therefore, in order to train the network more effectively, background patches are selected from more challenging locations, according to the following strategy, adapted from [44]. At every iteration, the network in its current state is applied to the 32 B-scans from which the fovea patches are extracted. From these 32 B-scans, the misclassifications of the background pixels by the current network are used to identify challenging locations. Each background pixel is assigned a weight $w_i = |y_i - l_i|$. Here, y_i equals the classification of the current network for pixels i, and l_i its label. The weights are now used to draw a weighted random sample at 32 locations, where the probability for each background pixel to be included is calculated as:

$$p_i = \frac{w_i}{\sum_{j \in X_-} w_j} \tag{2}$$

Here p_i is the probability of background pixel i to be included in the sample and X_- is the set of all background pixels in the current set of 32 B-scans. In this way, misclassified background pixels are more likely to be used during training. For computational efficiency, background pixels from other B-scans than the one with the fovea annotation are not included in the training procedure. Although background pixels from other B-scans may provide some additional information to train the network, we suspect that the foveal B-scans contain sufficient challenging samples, due to e.g. presence of fluid, or in some cases the optic disk.

In order to artificially increase the variation in the training set the patches are augmented by applying horizontal flipping with a 50% chance, and a random rotation between $-10°$ and $10°$. The network is trained for 16 epochs in total: 8 epochs with learning rate 0.001 and 8 epochs with learning rate 0.0001. Here, one epoch is defined as a pass over all data, consisting of 55 iterations, where 32 of the 1744 OCT volumes are used per iteration.

3.3. Classification

Although the network is only trained on B-scans containing the foveal center, it is applied to all B-scans in an OCT volume to find the foveal center in unseen scans, where the B-scans are zero-padded to compensate for the loss of border pixels. In this way a likelihood of belonging to

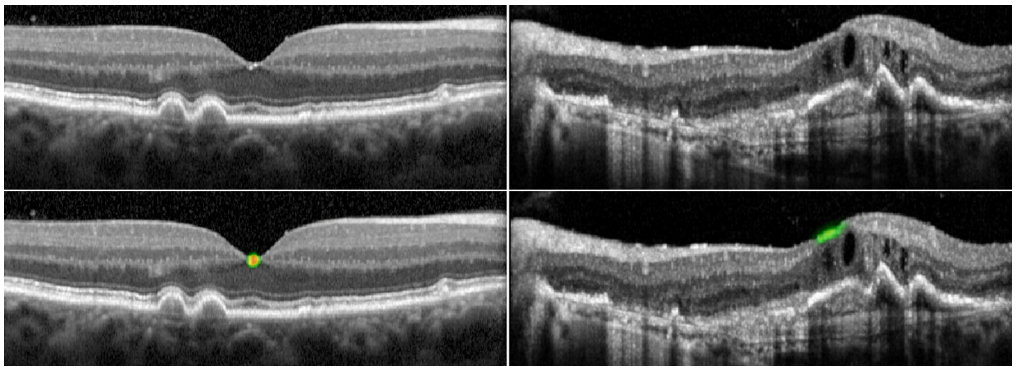

Fig. 4. Preview of the likelihood maps as generated by the network. Top images show the input foveal B-scan. The bottom image includes a heatmap overlay representing the fovea likelihood. The left example represents an intermediate AMD case, whereas the right case represents advanced AMD.

the fovea for each pixel in the volume is obtained. Next, the probability volume is smoothed with an anisotropic three dimensional Gaussian filter. The Gaussian smoothing could remove possible outliers and may improve accuracy. Finally, the pixel with the global maximum probability after smoothing is chosen as the location of the fovea. Figure 4 shows a preview of the fovea likelihood map for the foveal B-scan for an intermediate AMD and an advances AMD case.

4. Experimental design

The predicted fovea location of the proposed automatic method is compared with the reference standard. We only take into account the lateral and transversal distance, because clinically the axial position of the fovea is not well defined, nor is it very relevant for our application. Moreover, deviations from the reference location in the axial direction are expected to be small compared to the other two dimensions, due to the high axial resolution of 3.9 µm.

To derive other statistics, such as the mean distance to the reference, we discard detections with a distance to the reference location larger than the anatomical foveal radius, i.e. 750 µm [45]. This large distance is chosen even though the radius of the clinical fovea is much smaller (250 µm). The reason for this is that for some OCT volumes the distance between B-scans is larger than 250 µm. Therefore, if we would cutoff at 250 µm, we may discard detections in adjacent B-scans that may be inaccurate, but not necessarily wrong. We then compare the performance of the proposed method with the two independent human graders and the scan center as foveal center.

To evaluate the performance of the proposed method on the external data set, the CNN will be retrained with settings identical to the proposed method, using just the 40 scans that were selected for training. Because this training set is much smaller than the original training set, the same scans will be used multiple times per epoch.

To find out what distinctive features of the proposed method are essential for accurate detection of the foveal center, we perform additional experiments with variations of the proposed architecture and variations of the proposed training procedure. See Table 2 for an overview of the alternative network architectures that are used. For each variation the same training procedure is used, as described above. We will refer to the proposed network architecture as network A. As a first variation, we test a network without the dilated convolutions (network B). This network architecture is identical to network A, except that each dilated convolution layer is replaced with a regular 3x3 convolution. As a consequence, the receptive field of this network is much smaller (only 17×17 pixels). We also test a network without the shortcut connection (network C). This network again has the same architecture as network A, but without the concatenation

layer. Instead, convolution layer 9 is directly connected to the last dilated convolution layer (layer 7). Next, we test two modifications to the size of the dilation factors in the layers with dilated convolutions. Here, we use isotropic values for the horizontal and vertical dilation factors instead of the anisotropic values proposed in Table 1. Network D uses isotropic dilation factors of (2, 4, 8, 16, 32) for layers 3 to 7, respectively, and, consequently, has a receptive field of 131×131 pixels. Network E uses dilation factors of (3, 6, 12, 24, 48) and, consequently, has a receptive field of 193×193 pixels. In addition to these variations we test a different, more traditional CNN architecture, based on VGG [46]. This network contains five blocks of two 3×3 convolutional layers, with a max-pool layer in between the blocks. The first layers use 32 filters per layer and after each max-pool operation the number of filters is doubled. The lost resolution due to the max-pooling operations is recovered using shift-and-stitch [40].

Table 2. Variations of the proposed network architecture.

Network	Dilated convolutions	Shortcut	Pooling	Receptive Field
A	yes, anisotropic	yes	no	193×131
B	no	yes	no	17×17
C	yes, anisotropic	no	no	193×131
D	yes, isotropic	yes	no	131×131
E	yes, isotropic	yes	no	193×193
F	no	no	yes	140×140

Two variations to the proposed dynamic training procedure are investigated, using the originally proposed network architecture (network A). First, we investigate the use of basic random sampling (RS) of background patches instead of the proposed dynamic sampling strategy. Because it may take longer for the network to converge if the background patches are sampled randomly [44], this network is trained 8 times longer than the proposed method: 128 epochs in total (64 epochs with learning rate 0.001 and 64 epochs with learning rate 0.0001). The second variation is to train fully convolutionally (FC) on a single B-scans per iteration. That is, every iteration the network will see the entire B-scan that contains the fovea. Classification errors are then back-propagated for the 5×5 pixels around the fovea and the background pixels outside the 50×50 pixels around the fovea. Similar to the proposed method, this network is trained for 16 epochs in total: 8 epochs of 1744 iterations with learning rate 0.001 and 8 epochs with learning rate 0.0001.

Finally, the optimal value for the σ of the Gaussian filter is estimated from classifying 400 OCT volumes in the training set with the final CNN, using the proposed method (network A). Here, candidate values for σ are drawn from $\{(i\sigma_a, j\sigma_l, k\sigma_t)|i, j, k \in [0, 1, ..., 4]\}$ where σ_a, σ_l and σ_t define the resolution of the grid of candidate σ-values in the axial, lateral and transversal direction, respectively. Based on observations from preliminary tests, we choose $\sigma_a = 1.25\,\mu m$, $\sigma_l = 5\,\mu m$ and $\sigma_t = 50\,\mu m$. Distances between the voxel with maximum fovea likelihood in the smoothed probability volume and the training reference location are calculated for all candidate σ values. Based on the distribution of these distances, the optimal value for σ will be determined.

5. Results

The proposed automatic method detected the fovea in 391 of the 400 (97.75%) OCT volumes in the test set with a distance to the reference observer smaller than 750 µm. The mean (± SD) distance to the reference for these volumes was 71 µm ± 107 µm. Of the 9 volumes with a larger distance, most occurred in the advanced AMD subset: 1 control, 3 early AMD, 0 intermediate AMD, 5 advanced AMD.

Figure 5 shows boxplots of the distances to the reference location for the proposed method, the two observers and the scan center for each AMD severity level. The two observers both annotated

399 of the 400 test cases with a distance smaller than 750 µm, with a mean (± SD) distance of 57 µm ± 84 µm and 56 µm ± 80 µm, respectively. The OCT scans in the test set were not always accurately centered on the fovea: in 21 of the 400 test cases (5.25%) the scan-center was not located within 750 µm of the reference annotation. The scans were poorly centered on the fovea especially in advanced AMD cases, where the distance to the reference was larger than 750 µm in 14 of the 100 cases (14%). The number of correct detections, with a distance of less than 250 µm (the clinical fovea), was 364 (91%) for the proposed method, compared to 379 (94.75%), 384 (96%) and 306 (76.5%) for observer 1, observer 2 and the scan center, respectively.

On the DME and CSC set, the fovea was found within 750 µm of the annotated location for 93 of the scans (93 %, 46 DME, 47 CSC), with a mean distance of 172 ± 132µm.

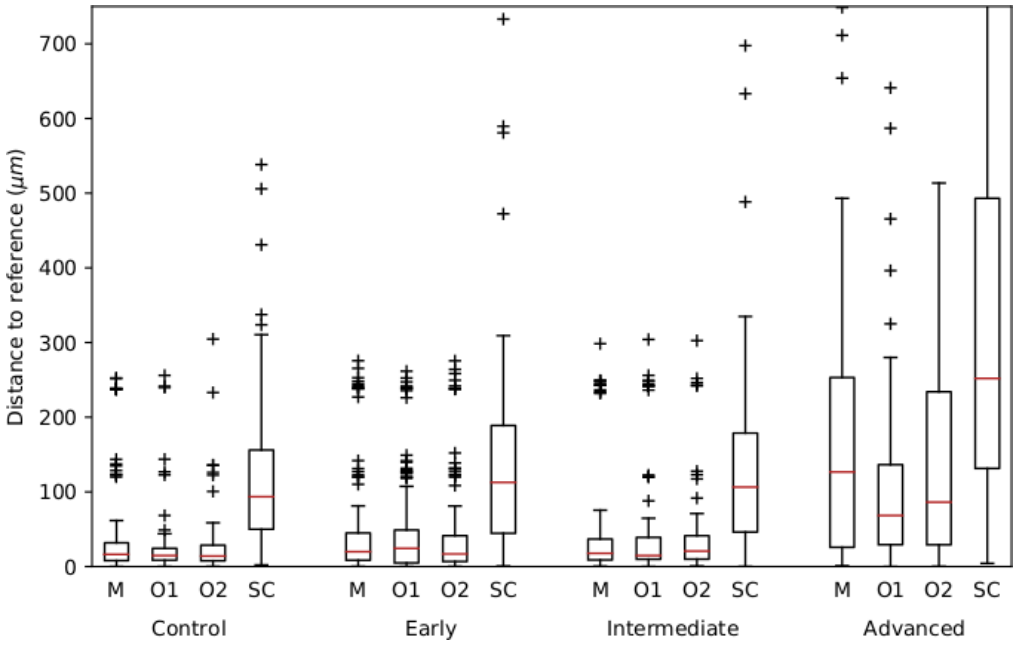

Fig. 5. Boxplots of the distance to the annotation per AMD severity level for the automatic method (M), the two observers (O1 and 02) and the scan center (SC). The errors (distance > 750 µm) are included for the creation of the boxplots, but are cut off from the figure for visualization purposes.

On the external data set, the proposed (retrained) method detected the fovea with a distance smaller than 750 µm in 328 of the 344 OCT scans (95.35%), with a mean (±SD) distance of 63 µm ± 84 µm. Again, most errors occurred in pathological retina's: the fovea was missed in 1 out of 95 control cases (1.05%) and in 15 of the 249 AMD cases (6.02%).

Table 3 summarizes the performances of the different network architectures. Network architecture B (with regular convolutions instead of dilated convolutions) shows a significant drop in performance, with 227 errors (56.75%). The other network architectures (C, D, E and F) show similar performance to the proposed method, with 9 (2.25%) to 11 (2.75%) errors. The two variations to the training procedure (RS and FC) have resulted in lower performance compared to the proposed dynamic method. The number of errors was 37 (9.25%) for RS and 30 (7.50%) for FC (see Table 4).

The grid search over possible values of the σ parameter of the Gaussian smoothing operation resulted in a σ of (3.75 µm, 15 µm, 100 µm) in the axial, lateral and transversal direction, respectively. When using the raw probabilities as predicted by the CNN, the fovea was missed in 11 of

Table 3. Results for different variations of the proposed network architecture. Values in the columns indicate the number of detections within the specified distance category (in μm).

Network architecture	<25	25-75	75-175	175-750	>750
A - Proposed method	215	72	44	60	9
B - No dilation	71	53	21	28	227
C - No shortcut	217	76	43	54	10
D - 131x131	225	63	43	58	11
E - 193x193	217	73	39	62	9
F - Pooling	219	61	46	63	11

Table 4. Results for the variations to the proposed training procedure. Number of updates refers to the number of iterations, or update operations (back-propagation) that were made during training. Training time refers to the total time needed to train the CNN.

Training procedure	Errors (>750 μm)	Number of updates	Training time
1 - Proposed method	9	880	6.58h
2 - Random sampling	37	7040	2.20h
3 - Fully convolutional	30	27904	8.18h

the 400 test cases (2.75%), which was reduced to 9 (2.25%) after the smoothing operation.

6. Discussion

The proposed method, using a fully CNN with anisotropic dilated convolutions that was trained using a dynamic training procedure, was able to detect the fovea in OCT volumes successfully. For healthy retinas as well as retinas affected by severe pathologies related to AMD, the accuracy is comparable to human observers. Figure 6 shows some examples of correct detections of the fovea. For a large subset of volumes the automatic method agreed almost perfectly with the human reference. The distance between reference and prediction was less than 1 pixel (11.5 μm) in 123 cases (30.75%) and less than 2 pixels (23 μm) in 206 cases (51.5 %). In some cases, the proposed method predicted the foveal center close to the reference, but in an adjacent B-scan. In Fig. 5 this is visible as clusters of values around ~120 μm and ~240 μm, the typical distances between B-scans in this data set. Note though that this also holds for the two observers O1 and O2. As can further be seen in Fig. 6, the proposed method performs well even in case of severely disrupting retinal structures such as large cysts, fibrosis or atrophy. Also in case of poor image quality the method was able to identify the fovea correctly.

Although the proposed method was trained only on healthy patients and patients affected by AMD, it generalizes well to other retinal pathologies. In both DME and CSC fluid inside the retina may cause severe structural disruptions, but the proposed method was still able to identify the fovea in 93% of the cases. It is not feasible to validate the performance of the method for every possible pathology. It should be noted though that no assumptions specific to AMD have been made in the design of the proposed method. We therefore expect the method to become more robust if training samples from other pathologies are added.

It is hard to make a fair comparison of the performance of the proposed method to other methods. Different methods were evaluated on different data sets with differences in image resolution, image quality, and type or severity of the pathology. Moreover, some performance metrics may refer to the subset of correct classifications, while others do take into account possible outliers. Table 5 includes a comparison of performance between methods, which should

be interpreted with care because of the differences in experimental setup between methods.

Table 5. Fovea detection performance reported by previous and current work. Accuracy (Acc.) refers to the number of detections within 750 µm of the reference annotation.

Method	N	Vendor	Pathology	Acc.	Distance (µm)	
Wang et al. [18]	100	Cirrus	Non-exudative AMD	N/A	104	± 62
Wu et al. [20]	80	multiple	Neovascular AMD	N/A	262.0	± 262.9
Montuoro et al. [25]	100	Cirrus	RVO	N/A	300	± 165 *
Proposed method	400	Spectralis	Control/AMD	97.75%	71	± 107
	100	Spectralis	Advanced AMD	95%	150	± 151
	50	Spectralis	DME	92%	215	± 140
	50	Spectralis	CSC	94%	130	± 109
	249	Bioptigen	Non-exudative AMD	94%	70	± 93

* Original values were reported in pixels instead of µm.

Figure 5 shows that the proposed method achieves a performance similar to the human observers. In one case however, there was a large disagreement between the observers and the reference standard. We suspect that the reference observer overlooked the true fovea here, while the method and the two human observers did correctly detect it (see Fig. 7). In other cases the agreement between human observers was good, and the mean distance between observer and reference was in line with previously reported values. We observed a mean (± SD) distance between reference and observers of 57 µm ± 84 µm and 56 µm ± 80 µm, respectively, while the authors in [17] reported 58.83 µm for OCT scans from Heidelberg Spectralis. The scan center is often a reasonable estimate of the true foveal center: only in 5.25% of the scans the distance between the scan center and the true fovea was larger than 750 µm. However, as can be seen in Fig. 5, it is not a very accurate measure. This is in line with previous work [17, 18] and corroborates the need for an accurate automatic method for detection of the foveal center.

In eight cases (ignoring the scan where the human observers did not agree) the automatic method failed to identify the fovea correctly and predicted a higher likelihood at a confounding location. Some examples of erroneous predictions can be found in Fig. 8 and Fig. 9. Two errors (control and early AMD) are due to a cut-off tilted retina, as in Fig. 8. In two early AMD cases, the network misclassified the fovea due to epiretinal membrane (ERM) formation (one in combination with an atypical fovea with ectopic inner foveal layers [47]). The remaining errors in the advanced AMD subset are due to formation of a confounding structure in combination with severe pathology near the fovea and/or poor scan quality, as in Fig. 9.

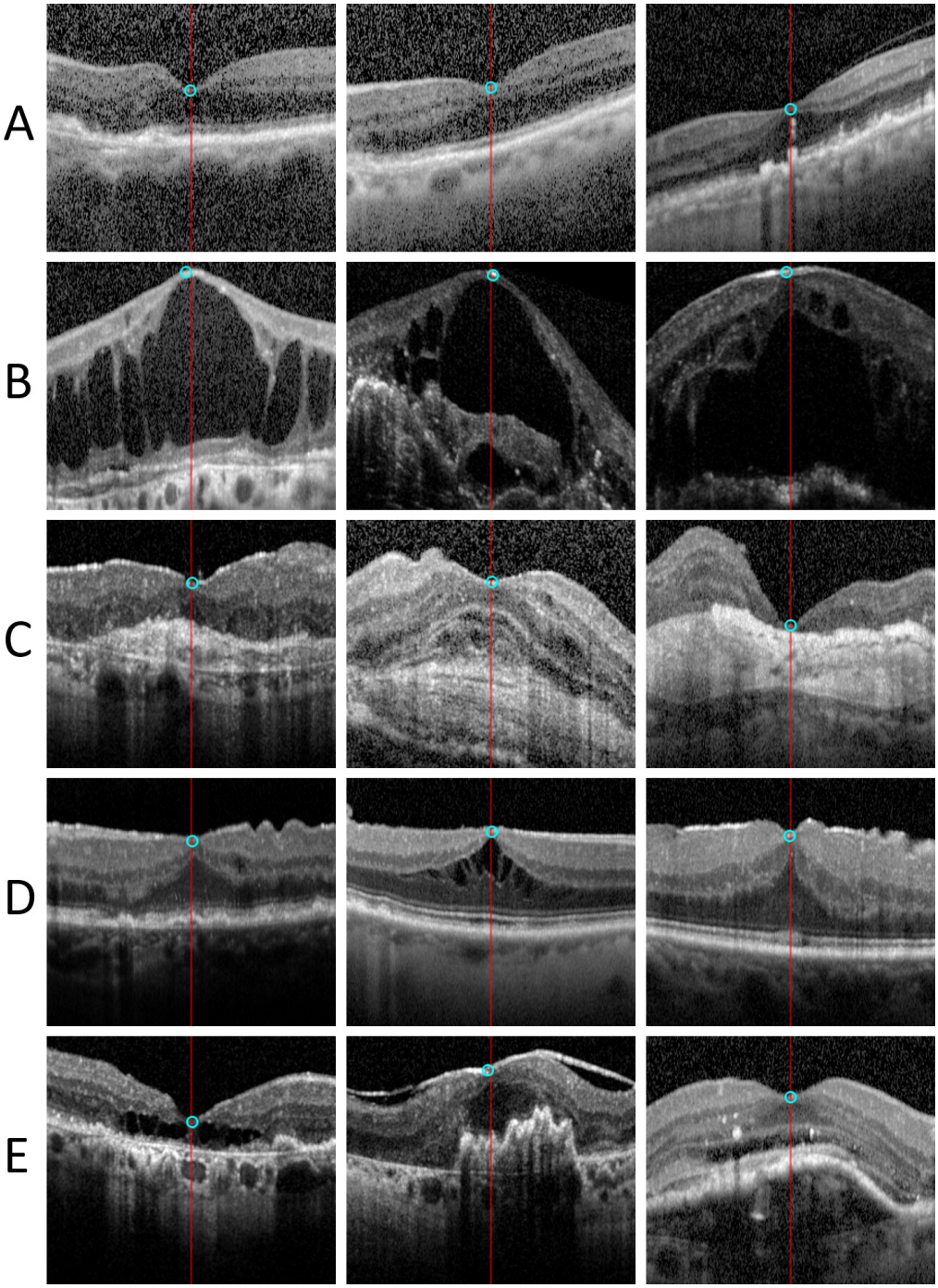

Fig. 6. Examples of correct detections of the foveal center (images are cropped and do not represent the full extend of the B-scan). The red line indicates the reference location, the cyan circle represents the predicted fovea location by the method. The method performs well even in case of A: noisy or tilted images; B: large cysts; C: structural disruption due to fibrosis; D: absent or minor foveal depression; E: other uncategorized structural deformations. All errors in these images are smaller than 62 µm.

For the external set only 40 scans were needed for training to obtain satisfactory results (95.3% accuracy). We believe it is possible to use such a low number of training samples because this data set has a more homogeneous nature than the EUGENDA data set: there are less scans of severely disrupted retinas because there is no advanced AMD and the quality of scans is much more comparable between scans. Adding more training data will likely improve the performance, whereas it should also be possible to mix the two data sets during training, to obtain a single network that can handle data from both vendors. When applying the proposed method (without retraining) on the external data set, performance was very poor with just 15.4% correct detections. The reason for this is twofold. First, due to the different nature of the scans (much more noise, less distinct layer boundaries), the response of the network at the fovea is much lower than for the scans from Heidelberg Spectralis. Second, most scans in the external data set have bright artifacts at the top of the image, with pixel intensities usually larger than average pixel intensity in the retina. In combination with the zero-padded border that is added to the top of the image, this may easily confuse the network. Therefore, in many scans the network gave a higher response at the top of the image than at the fovea. The solution that we proposed in this paper is to retrain the network on the external data set, demonstrating that it is feasible to apply our method to both data sets. Alternatively, a denoising algorithm [48, 49] could be applied to standardize input across vendors, which may result in better cross-vendor applicability.

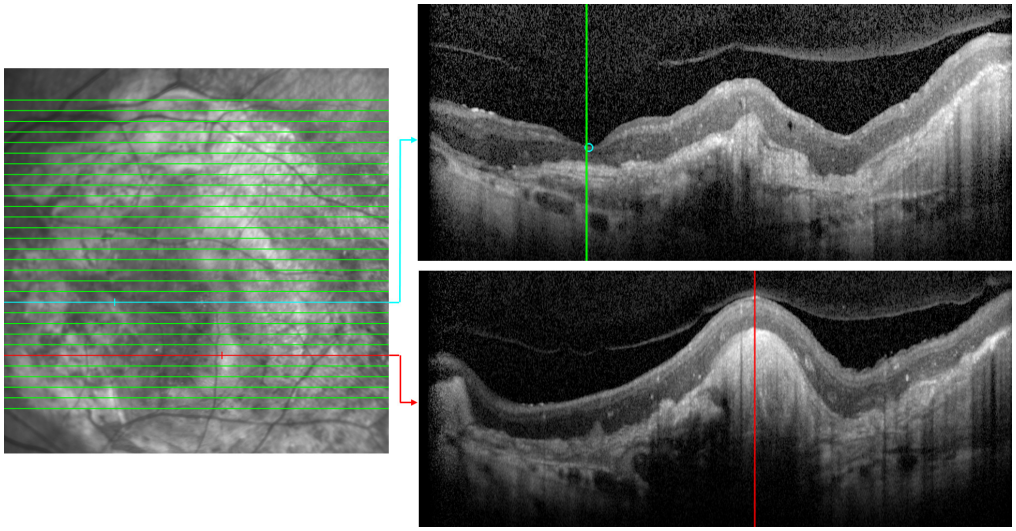

Fig. 7. Example image where the reference observer probably overlooked the true fovea location and annotated a confounding location in the retina. The red line indicates the reference location and the cyan circle is the predicted location of the automatic method. The (overlapping) green lines on the OCT indicate the locations predicted by the two observers.

From all variations to the network architecture, only architecture B showed significant decline in performance compared to the proposed method. The reason for this is most likely the small receptive field, because a contextual window of 17×17 pixels ($195.5 \times 66.3\,\mu$m) is usually not enough to reliably distinguish the fovea from other retinal structures. Unfortunately, the results from the experiments regarding the proposed innovations, such as the shortcut connection or the anisotropic size of the dilated filters, are inconclusive. The differences between the proposed architecture and the variations that were tested may be too small to notice significant differentiations. We do believe however, that the proposed innovations can be beneficial in other settings. By using anisotropic dilated convolution filters, we are able to tailor the size of the receptive field to our needs. This can be useful because the anisotropic nature of OCT

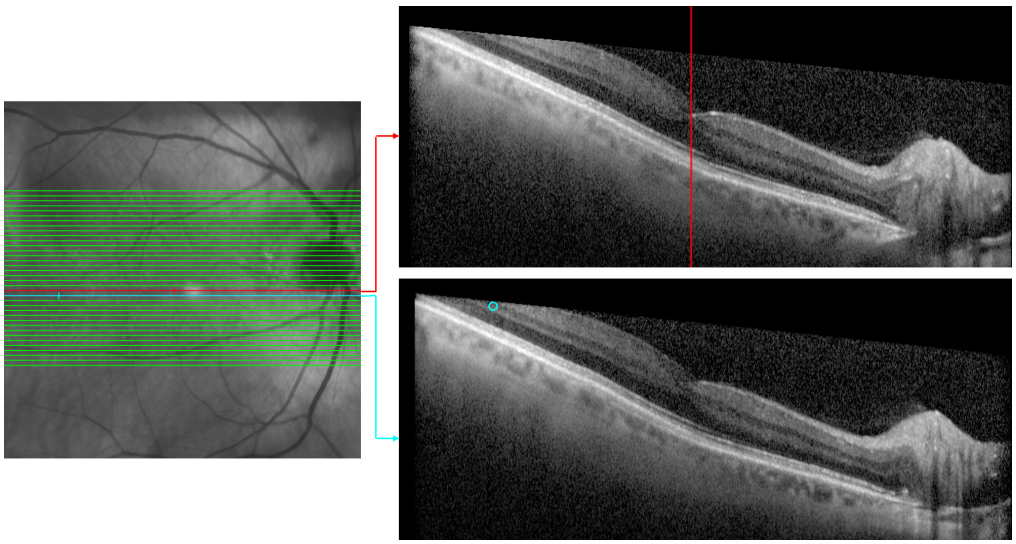

Fig. 8. Example error where the retina is tilted and cut-off at the top. The diagonal cut of the retinal layers to the right of the predicted location resembles a confluence of layers as typically observed near the fovea. In conjunction with the true fovea appearing a bit obfuscated, this has led to a misclassification.

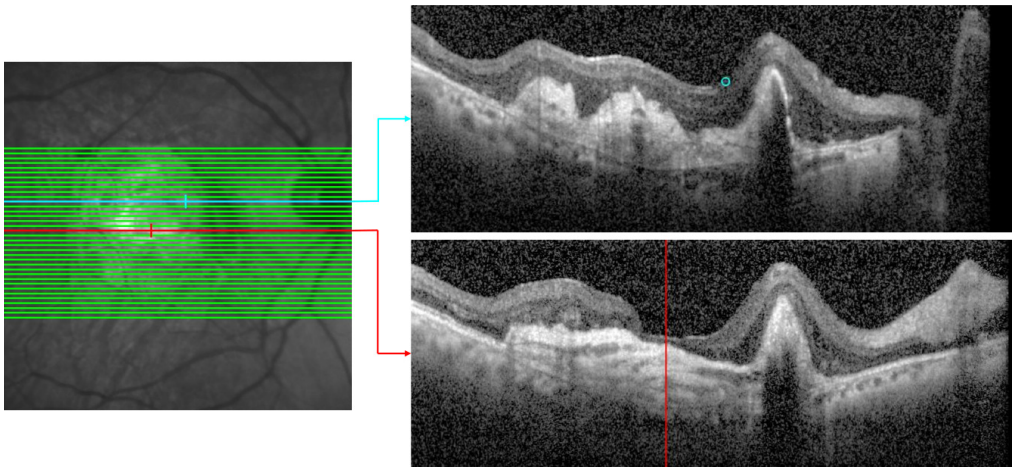

Fig. 9. Example error where the true fovea is affected and lacks many of the typical characteristics. The structure in the inner retinal layers disappears around the confounding location that was selected by the method. Therefore this location can easily be confused with the foveal center.

B-scans often demands a larger contextual window in the lateral direction. For this particular application, comparable performance can be obtained with the larger isotropic filters (network E). For other applications this may not be the case, as we may be limited by available memory or processing speed and if the receptive field is too large this could have an unfavorable impact on the pixel-accuracy. Finally, the design choices for the network architecture of the proposed method were made based on proposed architectures in literature, observations from preliminary test and intuition. The fact that the exact details of the network architecture are of minor influence on the final performance could be seen as a justification for the many ad-hoc decisions that are

often made in designing CNN architectures.

Network architecture F makes use of four max-pooling operations to increase the receptive field of the network. After every pooling operation the output resolution is halved, so the final resolution is $2^4 = 16$ times lower. To recover this we applied shift-and-stitch on the two-dimensional images, which means every B-scan had to be processed $16^2 = 256$ times. This is very inefficient, and the average classification time for a B-scan using shift-and-stitch is 3.18 seconds, compared to 1.31 seconds for the proposed method with dilated convolutions (measured on an Nvidia GTX 1080 GPU). In the current implementation, the method is applied to all pixels in every B-scan. Finding the foveal center can therefore take up to 48 seconds (37×1.31 seconds) per OCT volume. Arguably, this could be considered relatively inefficient, and for application to scans acquired with an even denser scanning protocol (e.g. 200 B-scans), this could become prohibitive. A solution would be to use a full retina segmentation [28], or an initial guess of the foveal center from a vessel mask obtained from the corresponding SLO image to exclude large regions of the background, which will make the implementation more efficient.

Two straightforward training procedures for training a fully CNN are random sampling of class-balanced patches and training fully convolutionally on larger patches (or even whole images). Both training procedures have been evaluated (RS and FC), but they demonstrated inferior results compared to the proposed dynamic sampling approach. We believe that for this application it will be very hard to obtain results as good as the dynamic sampling method with random sampling, even though it could be the case that some hyperparameters have not been set to optimal values for this experiment. The network will encounter difficult background patches only very rarely and this is most likely not enough to accommodate effective learning, also when the training time is prolonged even more.

The reason why training fully convolutionally (FC) produced inferior results remains unclear. The number of B-scans used to train the network using the FC training procedure was identical to the proposed dynamic sampling method. The difference is that the FC procedure makes 32 updates for 32 B-scans, while the proposed method only makes 1 update for a batch of 32 foveae and 32 selected hard background samples. Interestingly, the latter appears to be more effective. The proposed training procedure resulted in superior results even though less updates are made and the training data is only a sparse, albeit carefully selected, subset of all available data. It is possible that in order to obtain similar results with the FC training procedure we need to perform a more elaborate exploration of different hyperparameters, or possibly use a weight map for sampling the loss. These experiments lie outside the scope of this research.

6.1. Predictions at a sub-voxel level

Under the assumption that the true foveal center may lie anywhere in between B-scans, the expected distance from the foveal center to the closest B-scan is $\frac{1}{4}$ of the transversal resolution. The transversal resolution varies between images, but can get as large as 297 µm per pixel, which would mean an expected distance between the true foveal center and the closest B-scan of $\frac{297}{4} = 74.5$ µm. This distance is larger than the inter-observer agreement of about 57 µm, and is similar to the average distance error of the method. Hence, if we wish to obtain even more accurate predictions of the foveal center, it makes sense to allow for predictions at a sub-voxel level, i.e. in between B-scans.

A straightforward implementation that would allow for sub-voxel prediction would be to calculate the center of mass of all probabilities inside a bounding box around the pixel with maximum predicted probability:

$$\boldsymbol{c} = \frac{\sum_{i \in B} w_i \boldsymbol{r_i}}{\sum_{i \in B} w_i} \tag{3}$$

Here $\boldsymbol{c}$ is the predicted foveal center at sub-voxel level, B is the set of pixels inside the bounding box, w_i is the assigned probability of belonging to the fovea for voxel i, and $\boldsymbol{r_i}$ is the position of

voxel i (a three-dimensional vector).

Although we do not have human annotations at a sub-voxel level to validate whether the true foveal center is indeed better approximated by this method, we may get some indirect indications of the validity of the predictions. Assuming the errors of the human graders are random and unbiased, the mean of the annotations made by the three observers may be a more accurate prediction of the true foveal location than any individual prediction. Therefore, we expect that the mean distance between the center-of-mass method and the mean of the observers is smaller than the distance of the originally proposed method to the reference location.

Indeed this is the case, but for the 391 correct classifications the difference is not significant: 71 µm ± 107 µm originally, compared to 68 µm ± 69 µm for the center of mass to the mean of the observers ($p = 0.37$, two-sided T-test). It should be noted though that in case all observers annotated the foveal center in the same B-scan, the mean of the observers will still lie in this same plane even in cases the true foveal center may be slightly off. Therefore, we may need to look at the set of OCT scans where the three human observers differed in the B-scan in which they made the annotation (72 scans, excluding scans where the method failed). For these scans, there is a significant improvement of calculating the center of mass: 139 µm ± 123 µm mean distance for the original method to the reference, compared to 78 µm ± 73 µm for the center-of-mass to the mean of the observers ($p < 10^{-5}$). This could indicate that we can indeed achieve a more accurate prediction of the foveal center at a sub-voxel resolution.

To conclude, we have demonstrated how fully CNNs with anisotropic dilated convolutions can be applied to accurately detect the foveal center in an OCT volume. Performance of the proposed automatic method is comparable to human observers on a data set containing scans acquired with different scanning protocols, variable image quality, and presence of severely disrupting pathologies. Thus, the proposed method could facilitate automatic analysis of large macular data sets even in the presence of significant retinal pathology.

Funding

This work is funded by Radboud University, Radboud University Medical Center and the Fraunhofer-Gesellschaft as an ICON project.

Disclosures

The authors declare that there are no conflicts of interest related to this article.